Tijdschrift voor Onderwijsresearch, 20 (1995), nr. 1, pp. 34-41

De opbrengsten van het basisonderwijs: een repliek

W.J. van der Linden[1]
*Universiteit Twente**

M.A. Zwarts[1]
Inspectie van het Onderwijs

ABSTRACT

In this rejoinder to Wijnstra (1995), it is argued that, unlike the author seems to suggest, the definitions of the qualifications in the CEB evaluation project were not based on any assumptions regarding the distribution of difficulties in the item domains. Also, it is pointed out that statistical descriptions of the system of elementary education are contingent on the conditions under which the system exists, and that such descriptions are not sufficient to criticize the effects of the recommendations for change made by CEB. Finally, it is shown that the author's claim of aggregation of achievement scales being appropriate only for certain (unspecified) distributions of item difficulties ignores the unidentified character of the IRT difficulty scale, and hence does not apply.

Op 12 januari 1994 heeft de Commissie Evaluatie Basisonderwijs (CEB) zijn werkzaamheden voltooid met de aanbieding van een eindrapport en een viertal deelrapporten (Commissie Evaluatie Basisonderwijs, 1994a, 1994b, 1994c, 1994d, 1994e) aan de Minister van Onderwijs. Daarmee kwam een einde aan een intensieve werkperiode, die ruim twee jaar heeft geduurd. Het behoorde tot de opdracht aan de CEB om een zo breed mogelijke, evaluerende beschrijving van de toestand van het basisonderwijs te geven. Vervolgens moest deze evaluatie toegespitst worden aan de hand van een viertal specifieke vragen. Een volledig overzicht van de opdracht en de werkzaamheden van de CEB wordt gegeven in Janssens (1995).

Wijnstra (1995) heeft gelijk als hij in zijn notitie stelt dat de CEB bij zijn evaluatie van de opbrengsten van het onderwijs uitvoerig gebruik heeft gemaakt van de peilingen van de opbrengsten van het onderwijs door PPON. Op momenten als deze blijkt het belang van de beschikbaarheid van systematische empirische gegevens over de opbrengsten van het Nederlandse onderwijs. Gelukkig is het land dat goed peilingsonderzoek heeft.

Voor het betoog hieronder is het van belang om de precieze rol die de gegevens uit de PPON-projecten in de evaluatie van de CEB gespeeld hebben, scherp in het oog te houden. De CEB heeft de PPON-gegevens alleen gebruikt voor een *beschrijving* van de opbrengsten van het basisonderwijs. De *beoordeling* van deze opbrengsten is daarentegen volledig voor de rekening van de CEB. Bij de totstandkoming van deze beoordeling hebben ook deelbeschrijvingen uit andere bronnen voorgelegen (OVB-evaluatie; diverse SVO-projecten; dissertatie-onderzoek). Verder werd er door de CEB beoordeeld in het licht van evaluatieve uitspraken die eerder door anderen geuit waren. Sommige uitspraken zijn te vinden in de verslagen van de voornoemde projecten, terwijl andere werden geuit tijdens de raadpleging van zo'n 260 vertegenwoordigers van relevante instanties in het onderwijs en overige deskundigen. Ook werden in de PPON-projecten voor diverse leergebieden beoordelingsexperimenten uitgevoerd, waarin ouders, leraren BO en docenten VO hun verwachtingen en wensen ten aanzien van de opbrengsten kenbaar maakten. Naar het oordeel van de CEB was het algemene niveau van de verwachtingen en

[1] Beide auteurs waren respectievelijk als lid en stafmedewerker verbonden aan de Commissie Evaluatie Basisonderwijs.

* Adres: Universiteit Twente, Faculteit Toegepaste Onderwijskunde,. Postbus 217, 7500 AE Enschede.

Dit is het laatste artikel behorende bij het themanummer.

wensen onrealistisch hoog en waren ze in dit opzicht derhalve niet bruikbaar. Wel bruikbaar was het patroon in deze beoordelingen over de diverse leergebieden.

Ten slotte moet vermeld worden dat de CEB zijn beoordelingen heeft opgesteld op basis van een totaalbeeld van de toestand van het basisonderwijs dat onder andere berustte op een beschrijving van het aanbod, de leermaterialen, de pedagogisch-didactische aanpak, de leerlingen- en lerarenpopulatie alsmede van de fysieke, beleidsmatige en politieke omgeving van het onderwijs. De beschrijving van de opbrengsten vormde dus slechts een deel van dit totaalbeeld. Na de publikatie van de rapporten heeft zowel de pers als de politiek zich gestort op de constatering van de CEB dat het huidige aanbod overladen is. Voor de CEB waren andere constateringen, zoals die van het feit dat in het basisonderwijs, mede op grond van gebrek aan faciliteiten, de mogelijkheden van onderwijs op maat amper benut worden en dat er nog weinig systematisch aan kwaliteitszorg wordt gedaan, even belangrijk. Dergelijke constateringen vormen een onlosmakelijk geheel met de beoordelingen van de opbrengsten.

In zijn notitie maakt Wijnstra een aantal kanttekeningen bij de door de CEB gevolgde werkwijze, waar deze notitie een repliek op vormt. Zijn kanttekeningen laten zich groeperen rondom de volgende drie thema's:

1. de afhankelijkheid van de beoordelingen door de CEB van de toevallige psychometrische eigenschappen van de itemdomeinen;
2. de vraag of de normering door de CEB wel realistisch is geweest;
3. de problematiek van de aggregatie van de PPON-schalen.

ITEMDOMEINEN

In vrijwel alle peilingsonderzoek worden de opbrengsten van het onderwijs vastgesteld met behulp van itemdomeinen waaruit deelverzamelingen items aan verschillende steekproeven leerlingen worden voorgelegd. Het voordeel van deze werkwijze is dat de opbrengsten van het onderwijs met een maximale validiteit en nauwkeurigheid beschreven kunnen worden. Voorwaarde is wel dat de itemdomeinen omvangrijk en van een goede kwaliteit zijn. Om een hoge inhoudsvaliditeit te realiseren wordt een itemdomein in twee stappen geconstrueerd. In de eerste stap wordt het betreffende leer- of vormingsgebied volledig geclassificeerd. In de PPON-projecten staat deze stap bekend als de "domeinbeschrijving". In de tweede stap wordt een verzameling items geschreven die de gewenste verdeling over deze classificatie vertoont.

Het is de stellige indruk van de CEB dat zowel de domeinbeschrijvingen als de itemdomeinen uit de PPON-projecten een hoge inhoudsvaliditeit bezitten. Deze indruk is gebaseerd op kennisname van het materiaal en de lof die inhoudsdeskundigen aan PPON hebben toegezwaaid. Het resultaat is het gevolg van een werkwijze waarin bij de constructie van de opgaven de inhoudsvaliditeit voorop stond (Wijnstra, 1988, pag. 5). Uit commentaar van derden blijkt hoe de kwaliteit van de itemdomeinen voor bijvoorbeeld rekenen wordt gewaardeerd:

> "De PPON-rekentoetsen zijn een uitdrukking van een range van eindtermen: kennis en vaardigheden waarvan te verdedigen valt dat alle kinderen aan het einde van de basisschool ze zouden moeten beheersen' (Teunissen, 1988, pag. 169).
>
> "Nadere bestudering van de voorbeeldopgaven en van de totale collectie wijst uit dat de PPON-toetsen inderdaad het overgrote deel van de eindtermen bestrijken. ... Voorts kan worden vastgesteld dat de kwaliteit van de PPON-opgaven in het algemeen ruimschoots voldoende is om het onderwijsniveau passend te peilen" (Treffers, 1988, pag. 182).

Alleen Huitema (1988, pag. 164) maakt een opmerking waarin hij aangeeft dat hij meer opgaven had willen zien die minder van de Nederlandse taal afhankelijk zouden zijn en daardoor gemakkelijker waren. In het onderzoek van Zwarts en Janssens (1994), waaraan Wijnstra in de inleiding van zijn kanttekeningen refereert, gaf het panel van beoordelaars aan dat de toetsen voor rekenen goed samenvallen met de kerndoelen. Dezelfde conclusie werd getrokken voor Nederlandse taal, Engelse taal en verkeer. Alleen voor enkele domeinen van wereldoriëntatie moet er een voorbehoud worden gemaakt.

Naast de inhoudsvaliditeit wordt een tweede kenmerk van itemdomeinen gevormd door hun verdeling van de psychometrische eigenschappen van de items. Het is op dit punt dat de notitie van Wijnstra zich richt. Alvorens we de precieze inhoud van zijn kanttekening aangeven, is het van belang om op te merken dat in de PPON-projecten niet bewust naar een bepaalde verdeling van de psychometrische eigenschappen werd gestreefd. Dit gegeven wordt door Wijnstra verwoord wanneer hij stelt dat:

> "Door deze schalingsprocedure is er geen pertinente noodzaak te streven naar een bepaalde gemiddelde moeilijkheidsgraad of een bepaalde verdeling van de p-waarden van de opgaven en *dat is ook nooit gebeurd*" (pag. 4).

Zulk een streven zou in de praktijk ook nooit gerealiseerd hebben kunnen worden. Talloze empirische onderzoekingen laten namelijk zien dat, met uitzondering van enkele experimenten met matig succes voor onderdelen uit de wiskunde, het vrijwel onmogelijk is om verschillen tussen de eigenschappen van items die bij eenzelfde "cel" in een domeinbeschrijving behoren te voorspellen of te verklaren. Zulke items, die alle geschreven zijn bij eenzelfde onderwerp, op hetzelfde gedragsniveau en voor dezelfde leerlingen, blijken psychometrische verschillen op de leveren die in hoge mate afhangen van "toevallige", ogenschijnlijk irrelevante eigenschappen van de formulering van de items. De enige wijze waarop een gewenste verdeling van de psychometrische eigenschappen van de items kan worden verkregen is achteraf, door selectie aan de hand van empirische schattingen van deze eigenschappen. Wijnstra refereert aan deze bevindingen wanneer hij het volgende neerschrijft:

> "Het is een bekend gegeven dat het beoordelen van de moeilijkheidsgraad van opgaven een notoir moeilijke zaak is" (pag. 5).

We kunnen nu het door Wijnstra opgeworpen "meningsverschil" met de CEB scherp stellen. De CEB heeft, overeenkomstig de werkwijze van PPON, aangenomen dat de feitelijke verdeling van psychometrische eigenschappen in de itemdomeinen allereerst bepaald is door de opgestelde domeinbeschrijvingen en de gekozen verdeling van de items over de cellen van deze domeinbeschrijving. Gezien het toevalskarakter van de overblijvende verschillen tussen de items en de grootte van de geaggregeerde itemdomeinen waar de CEB mee heeft gewerkt, is vervolgens een beroep gedaan op de wet van de grote aantallen en werd aangenomen dat de feitelijk gecreëerde itemdomeinen psychometrisch een afspiegeling vormen van de domeinen aan mogelijk items die men zich bij de domeinbeschrijvingen in kan denken. Onder deze aanname is een benadering mogelijk die bekend staat als "domain-referenced measurement" en waarbij normen geformuleerd kunnen worden in termen van de aantal-goed scores van de populatie leerlingen voor het totale domein (zie bijvoorbeeld Hively, Maxwell, Rabehl, Sension & Lundin, 1973). De CEB heeft deze benadering gekozen; voor een definitie van de normen raadplege men Van der Linden en Zwarts (1995). Deze normering is uiteraard niet blind geweest en berust op een inhoudelijke beoordeling van alle beschikbare PPON publikaties en materialen. Voor een tweetal onderdelen van rekenen en taal alsmede voor het gehele gebied van wereldoriëntatie, bestond de overtuiging dat de hiervoor omschreven aanname niet volledig correct was en is door de CEB een correctie op de normen toegepast.

Het merkwaardige is nu dat Wijnstra de hiervoor beschreven werkwijze in de PPON-projecten nog eens herbevestigt, maar vervolgens suggereert dat de normering van de CEB alleen maar zin heeft als er binnen PPON *wel* naar een vooraf bepaalde verdeling van de moeilijkheden van de items in de itemdomeinen zou zijn gestreefd. Aan welke voorwaarden voldaan zou moeten worden en hoe deze "bepaalde" verdeling van itemmoeilijkheden er uit zou moeten zien, wordt evenwel niet meegedeeld. Op dit punt moet er echt sprake zijn van een misverstand. De werkwijze van de CEB berust niet op een aanname dat alle PPON-items een p-waarde in de buurt van .70 zouden bezitten of iets dergelijks. Zoals hierboven geschetst werd, is het enige dat gevraagd wordt dat de itemdomeinen representatief zijn voor de leerstof.

Het is wel zo dat men in tweede instantie, door itemanalyse op grond van empirische gegevens, de samenstelling van de oorspronkelijke domeinen kan wijzigen. Door PPON is langs deze weg een aantal items uit de oorspronkelijke domeinen verwijderd. Dit aantal was telkens klein. Het doel bij deze verwijdering was om schalen te creëren met een maximale breedte aan

waarden voor de moeilijkheidsparameter uit het gehanteerde IRT model. Zo werden bijvoorbeeld enkele items verwijderd die redundant waren omdat ze deel uitmaakten van een cluster van items met ongeveer dezelfde parameterwaarden. Ondanks de kleine aantallen waarom het gaat, voor rekenen bedraagt het aantal zo'n 4% van het oorspronkelijke domein, zou deze itemselectie de beoordelingen van de CEB hebben kunnen verstoren. Daarom is door de CEB voor alle zekerheid een robuustheidsanalyse uitgevoerd waarin deze itemselectie werd gesimuleerd en de gevolgen nagegaan werden. Voor een volledig overzicht van de resultaten van deze analyse wordt verwezen naar de figuren 3-6 in Van der Linden en Zwarts (1995). Het vlakke verloop van bijna al deze figuren over de gehele range van 0-100% wijst op een extreem hoge robuustheid van de beoordeling van de CEB ten opzichte van het gevolgde type itemanalyse voor de itemdomeinen in de PPON-projecten. Voor een groot aantal leergebieden moeten bijna alle items verwijderd worden alvorens het oordeel van de CEB zich dient te wijzigen. Alleen in een paar gevallen, waarin de gemiddelde geobserveerde scores voor de populatie leerlingen reeds dicht tegen de normen van de CEB aanzaten, kan er voor kleinere percentages verwijderde items een verandering van de beoordeling nodig zijn, die soms voor grotere percentages weer ongedaan gemaakt moet worden.

REALISTISCHE NORMERING?

Wijnstra vraagt zich hardop af wat de realiteitswaarde is van de door de CEB toegepaste normering. Terecht meet hij de betekenis van kwalificaties als "voldoende", "matig" en "onvoldoende" af aan de inspanning in het onderwijs die geleverd moet worden om de geobserveerde scoreverdeling naar een hogere kwalificatie op te laten schuiven. Deze operationele definitie was ook degene van de CEB en is er de reden van dat, zoals eerder reeds werd opgemerkt, de aanbevolen maatregelen en de beoordelingen niet los van elkaar gezien mogen worden. De kern van de opmerkingen van Wijnstra is dat de beoordelingen van de CEB effecten van maatregelen veronderstellen die zelden gevonden worden bij door het onderwijs manipuleerbare variabelen.

Met deze effecten en variabelen doelt Wijnstra ongetwijfeld op de door PPON gepubliceerde resultaten van analyses waarin de gemiddelde scores op de PPON-schalen gerelateerd worden aan achtergrondvariabelen als de gehanteerde leerboeken, geslacht, leeftijd, gewicht van de leerlingen in de formatiebepaling, e.d. De constatering dat in PPON effecten van de orde van grootte van 0,5 standaarddeviatie zelden geschat konden worden, is terecht. Voor bijvoorbeeld rekenen worden effecten van deze grootte vrijwel alleen aangetroffen, maar dan onmiddellijk voor een ongewenst groot aantal PPON-schalen, voor de gewichtscategorieën voor leerlingen zoals deze in de huidige financiering van het basisonderwijs gelden.

Door de in PPON gevonden effecten als norm voor onderwijsverandering te kiezen, maakt Wijnstra een klassieke fout in de interpretatie van de resultaten van statistische resultaten. Regressie van de resultaten van peilingsonderzoek op achtergrondvariabelen levert alleen een *beschrijving* op van de samenhang tussen variabelen onder de *huidige voorwaarden* in het onderwijs. De gedachte dat men daarmee causale verbanden vastgesteld heeft, die de effecten van onderwijsverandering voorspelbaar maken, is onterecht. Dergelijke verbanden kunnen nooit via peilingsonderzoek vastgesteld worden. Technisch gezien gaat het hier om het onderscheid tussen regressievergelijkingen die worden geschat in een beschrijvend onderzoek en differentiaalvergelijkingen die worden geverifieerd via experimenteel onderzoek. Tegen overspannen verwachtingen van het gebruik van regressie-analyse in peilingsonderzoek, werd reeds gewaarschuwd door Freudenthal (1975) in zijn bekende kritiek op de eerste internationale peilingen door de IEA, die in zijn rapporten dezelfde status aan regressie-analyse toeschreef.

In de tweede plaats is er genoeg empirisch onderzoek dat laat zien dat effecten van instructie- en schoolvariabelen op de leerprestaties groter dan 0,5 standaarddeviatie wel degelijk tot de mogelijkheden behoren. In een recent overzicht van Fraser (1989, Table 2.1) heeft maar liefst een derde van de bestudeerde variabelen een effectgrootte die boven deze norm ligt.

Er moet dus rekening mee gehouden worden dat de CEB zich in zijn aanbevelingen op een

grondige herziening van een aantal voorwaarden in het onderwijs richt. Zoals al gememoreerd werd, is een behoorlijke inperking van het verplichte aanbod één van de aanbevelingen, zodat er per saldo veel meer tijd voor minder onderwerpen zal ontstaan. Nu blijken te grote percentages scholen niet eens aan de behandeling van sommige onderwerpen toe te komen. Ook het vrijmaken van extra middelen om tijdens de eerste vier jaren van het basisonderwijs vroegtijdige diagnose te kunnen vervolgen met een aanbod aan onderwijs op maat, is een krachtige maatregel. Voeg daarbij nog eens het pleidooi voor een opbrengstgerichte cultuur, waarin kwaliteitszorg een vast gegeven is, dan is het niet overdreven om een pittige verhoging van de opbrengsten te verwachten. Om deze voorspelling te concretiseren het volgende. De curve voor Proportions/ Percentages in figuur 1 in Van der Linden en Zwarts (1995) heeft geleid tot het oordeel "onvoldoende". Het is helemaal niet onrealistisch om aan te nemen dat, waar nu zelfs gemiddeld zo'n 8-9% van de leerlingen überhaupt geen instructie in deze onderwerpen krijgt (Wijnstra, 1988, par. 5.7), de bovengenoemde maatregelen de curve een vorm aan kunnen doen nemen die dicht bij die van de eerste curve in deze figuur ligt, die de beoordeling "matig" vertegenwoordigt.

Overigens komen de beoordelingen van de CEB, op een enkele uitzondering na, vrijwel perfect overeen met die van het PPON-team in een recente aflevering van een themanummer van het blad *School* (1994). Dit themanummer verscheen gelijktijdig met de rapporten van de CEB en is onafhankelijk tot stand gekomen. Hieronder volgt een selectie van de evaluatieve uitspraken door het PPON-team, waarvoor we de lezer oproepen om deze met de beoordelingen in de CEB-rapporten te vergelijken:

"Als goed uit de bus komen de prestaties van leerlingen bij spellen en technische lezen. Als zwak: begrijpend lezen, schrijven en spreken." (pag. 24-25)

"Het is waar dat de kwaliteit van het schriftelijk werk in groep 8 soms erg tegenvalt. Het wonderlijke is dat we in groep 5 vergelijkbare opdrachten hebben gegeven, die goed werden uitgevoerd." (pag. 25)

"De resultaten op het onderdeel cijferen zijn goed te noemen: veel leerlingen beheersen de standaardalgoritmen voor de bewerkingen optellen, aftrekken, vermenigvuldigen en delen met gehele getallen." (pag. 28)

"De resultaten op het onderdeel procenten vallen tegen. Procentopgaven worden slechts door een kleine groep leerlingen goed beheerst. Een gemiddelde leerling heeft geen goed begrip van wat procenten zijn en hij is niet in staat eenvoudige procentberekeningen met een grote kans op succes uit te voeren." (pag. 29)

"Hoofdrekenen moet beter, schatten is onvoldoende, maar dat is ook nog vrij nieuw in het onderwijs. Inzichten in het fundament van breuken, verhoudingen en procenten moet veel beter. De basiskennis van procenten blijkt ook onvoldoende. In sommige methodes wordt er pas in groep 8 aandacht aan besteed. Er is ruimte om die lacunes te bestrijden, we cijferen immers te veel." (pag. 30)

"Bij taal vond ik het zeer opvallend dat de produktieve vaardigheden, stellen en spreken, zwaar beneden de verwachtingen bleven." (pag. 39)

"Spellen bleek boven verwachting goed." (pag. 39)

"Bij rekenen is het algemene beeld dat breuken, procenten en verhoudingen er mager uitkomen." (pag. 39)

AGGREGATIE VAN SCHALEN

In de PPON-projecten wordt gerapporteerd op een hoog niveau van detaillering. Bijvoorbeeld voor de opbrengsten van het rekenen aan het einde van het basisonderwijs werden de opbrengsten gepeild met behulp van 27 verschillende schalen. Het aantal schalen is mede het resultaat van het gebruik van itemresponstheorie (IRT), dat aan de schaling van de gebruikte opgaven de eis van eendimensionaliteit oplegt. Dit is een duidelijk winstpunt. We weten daardoor welke delen van leergebieden we als homogeen kunnen beschouwen en kunnen bijvoorbeeld bij de herziening van didactische methoden met dit gegeven rekening houden.

Wie naar het beleid rapporteert, is echter onderhevig aan de wetten die daarbij behoren. Je betreedt dan de wereld van de *executive summaries,* moet aansluiten bij de "lopende discussies" en vooral "heldere boodschappen" brengen. Voor de rapportage van de opbrengsten van het basisonderwijs ligt dat niet anders. Een beleidsmaker heeft er niet zoveel aan om te weten of het nu met "vermenigvuldigen" of met "delen" fout zit. Hij is meer geïnteresseerd in de opbrengsten voor de "basisbewerkingen" en wil antwoord op de vraag of daar in het onderwijs nu meer of minder aandacht aan geschonken moet worden dan aan bijvoorbeeld "toepassingen". Kortom, er is aggregatie van de gegevens nodig tot het niveau van de grotere onderdelen van de leergebieden–ongeveer het niveau dat gebruikt werd in de uitspraken van het PPON-team in de citaten hierboven. Dat is niet erg. Er gaat wel wat informatie verloren, maar er wordt aan effect gewonnen. En voor achtergrondinformatie of in discussies met vakgenoten kan altijd weer teruggegaan worden naar het oorspronkelijke niveau aan detaillering.

In de CEB-rapporten is welbewust gekozen voor een indelingsniveau dat relevant is voor de maatschappelijke discussie over de invoering van de kerndoelen in het basisonderwijs. Deze keuze brengt onherroepelijk met zich mee dat voor de scores een andere schaal gekozen moet worden dan de IRT-schalen die in de PPON-projecten werden gehanteerd. Door de CEB zijn daarom alle berekeningen gedaan op de schaal van de aantal-goed scores voor geaggregeerde itemdomeinen uit de PPON-projecten. Voor de berekening van de populatieverdelingen op deze schaal is dankbaar gebruik gemaakt van de schatting van de IRT-parameters van de items uit de PPON-projecten. Voorbeelden van deze berekeningen zijn te vinden in Van der Linden en Zwarts (1995).

Wijnstra maakt twee opmerkingen bij de gevolgde aggregatieprocedure. Allereerst stelt hij de vraag of een ongeveer gelijke weging van de afzonderlijke PPON-schalen wel terecht is en de opbrengsten op de schalen voor basiskennis en -begrip niet zwaarder gewogen zouden moeten worden dan de opbrengsten op de meer toepassingsgerichte schalen. Ten tweede presenteert hij in tabel 2 de verdelingen van de waarden voor de moeilijkheidsparameter uit de IRT voor de onderdelen procenten en verhoudingen en vraagt zich af "of de commissie hiermee bij de aggregatie niet op de een of andere manier rekening had moeten houden" (pag. 7).

Op de eerste vraag is onze reactie kort. De feitelijk gehanteerde weging verdient de voorkeur. De verdeling van de aantal-goed scores op geaggregeerde domeinen leidt tot totaaloordelen aan de hand van een omnibusmaat. Bij ongelijke weging zouden deze totaaloordelen zeer gevoelig worden voor de opbrengsten op slechts enkele van de oorspronkelijke schalen en zouden de opbrengsten op de andere schalen dus noodzakelijk delen in het oordeel dat de eerste afdwingen. Bij gelijke weging kan dezelfde gevoeligheidsproblematiek reeds spelen als één van de onderliggende vaardigheden of kennisgebieden in sterke mate van de andere afwijkt. Ook dit is ongewenst, maar onvermijdelijk. In zo'n geval heeft de CEB in zijn rapporten de toelichting op de tabellen gebruikt om deze afhankelijkheid zichtbaar te maken. Zie hiervoor bijvoorbeeld de toelichting op de rol die het lezen van rapporterende en argumentatieve teksten speelt in het totaaloordeel over de opbrengsten voor begrijpend lezen (Commissie Evaluatie Basisonderwijs, 1994a, pag. 58).

De reactie op de tweede vraag is methodologisch van aard. Vergelijking van de spreidingen van de twee verdelingen van de moeilijkheidsparameters voor procenten en verhoudingen in tabel 2, zoals Wijnstra doet, is zinloos. In feite worden hier appels met peren vergeleken. Onder het in PPON gebruikte IRT model hebben deze verdelingen namelijk ieder een willekeurige eenheid en nulpunt (zie bijvoorbeeld Hambleton & Swaminathan, 1985, par. 4.2). We kunnen de spreiding van de parameterwaarden op de individuele PPON-schalen zo breed of smal maken als we zelf willen, zonder hierdoor de passing van het model aan te tasten. Ieder van deze spreidingen zou ook *dezelfde* verdeling van de aantal-goed scores voor het betreffende itemdomein opleveren en derhalve tot *hetzelfde* oordeel van de CEB hebben geleid. Op de door Wijnstra opgeworpen vraag bestaat geen antwoord, omdat ze geen probleem aansnijdt.

DISCUSSIE

In deze reactie is betoogd dat er geen meningsverschil bestaat over de wijze waarop in de PPON-projecten de itemdomeinen zijn geconstrueerd. Volgens Wijnstra is alleen gestreefd naar een inhoudelijke dekking en niet naar een bepaalde verdeling van psychometrische eigenschappen van de items, en dat is precies de veronderstelling waaruit de CEB vertrokken is. Merkwaardig blijft het betoog van Wijnstra dat de werkwijze en normering van de CEB alleen zin hebben als de moeilijkheid van de items een bepaalde verdeling zou bezitten, waarvan hij de aard overigens niet omschrijft. De verwijzingen van Wijnstra naar de regressie van de PPON-resultaten op een aantal achtergrondvariabelen en de betekenis van de verdeling van de moeilijkheidsparameters in de itemdomeinen berusten beide op een methodologisch misverstand, waarvan de aard door ons is aangegeven. Blijft staan de vraag of de combinatie van beoordelingen en aanbevelingen door de CEB realistisch is geweest. Het deed ons deugd te constateren dat het PPON-team in een onafhankelijke publikatie deze beoordelingen vrijmoedig bijviel.

Het was niet de bedoeling van deze reactie om de betekenis van peilingsonderzoek voor het Nederlandse onderwijs ter discussie te stellen. Deze staat buiten kijf. In de aanhef is al gesteld dat op momenten als de evaluatie door de CEB het belang van systematisch empirisch onderzoek naar de opbrengsten van het onderwijs blijkt. In zijn nabeschouwingen heeft de commissie het belang van peilingsonderzoek onderstreept en ervoor gepleit om de systematiek van peilingsonderzoek uit te breiden naar een periodieke verzameling van empirische gegevens over aanbod, context en processen in het onderwijs. Ook is een systematische vergelijking van peilingsresultaten in internationaal verband gewenst (Commissie Evaluatie Basisonderwijs, 1994e)

LITERATUUR

Commissie Evaluatie Basisonderwijs (1994a). *Inhoud en opbrengsten van het basisonderwijs* (Deelrapport 1). De Meern: Inspectie van het Onderwijs.

Commissie Evaluatie Basisonderwijs (1994b). *Onderwijs op maat* (Deelrapport 2). De Meern: Inspectie van het Onderwijs.

Commissie Evaluatie Basisonderwijs (1994c). *Onderwijs aan jonge kinderen* (Deelrapport 3). De Meern: Inspectie van het Onderwijs.

Commissie Evaluatie Basisonderwijs (1994d). *Onderwijs gericht op een multiculturele samenleving* (Deelrapport 4). De Meern: Inspectie van het Onderwijs.

Commissie Evaluatie Basisonderwijs (1994e). *Zicht op kwaliteit* (Eindrapport). De Meern: Inspectie van het Onderwijs.

Fraser, B.J. Research syntheses on school and instructional effectiveness. *International Journal of Educational Research, 13*, 707-720.

Freudenthal, H. (1975). Pupils' achievements internationally compared: The IEA. *Educational Studies in Mathematics, 6*, 127-186.

Hambleton, R.K. & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer-Nijhoff.

Hively, W., Maxwell, G., Rabehl, G., Sension, D. & Lundin, S. (1973). *Domain-referenced curriculum evaluation: A technical handbook and a case study from the MINNEMAST project*. Los Angeles, CA: Center for the Study of Evaluation, University of California.

Huitema, S. (1988). We overvragen de basisschool. In J.M. Wijnstra (red.), *Balans van het rekenonderwijs is de basisschool* (PPON-reeks nr. 1). Arnhem: Instituut voor toetsontwikkeling Cito.

Janssens, F.J.G. (1995). De evaluatie van het basisonderwijs door de CEB: aanpak en werkwijze. *Tijdschrift voor Onderwijsresearch, 20*, 3-12.

Van der Linden, W.J. & Zwarts, M.A. (1995). *Tijdschrift voor Onderwijsresearch, 20*, 13-27.

School (1994). *Een beeld van een basischool (themakatern). 22*, 19-40.

Teunissen, F. (1988). Een hoge norm. In Wijnstra, J.M. (red.), *Balans van het rekenonderwijs is de basisschool* (PPON-reeks nr. 1). Arnhem: Instituut voor toetsontwikkeling Cito.

Treffers, A. (1988). Over de merkbare invloed van onderwijsmethoden op leerprestaties. In Wijnstra, J.M. (red.), *Balans van het rekenonderwijs is de basisschool* (PPON-reeks nr. 1). Arnhem: Instituut voor toetsontwikkeling Cito.

Wijnstra, J.M. (red.) (1998). *Balans van het rekenonderwijs is de basisschool* (PPON-reeks nr. 1). Arnhem: Instituut voor toetsontwikkeling Cito.

Wijnstra, J.M. (1995). De opbrengsten van het basisonderwijs volgens de CEB: enkele kanttekeningen bij de gevolgde normeringsprocedure. *Tijdschrift voor Onderwijsresearch, 20*, 28-33.

Zwarts, M.A. & Janssens, F.J.G. (1994). *Meten van kerndoelen met PPON-toetsen*. De Meern: Inspectie van het Onderwijs.

Manuscript ontvangen 9-8-1994
Definitieve versie ontvangen 7-11-1994